Finite populations and simple random sampling

```
## Set this up for your own directory
imageDirectory <- "MyAssignmentDirectory/img"  # e.g. in current "./img"
dataDirectory <- "MyAssignmentDirectory/data"  # e.g. in current "./data"
path_concat <- function(path1, path2, sep="/") paste(path1, path2, sep = sep)
```

---

30 marks

Consider a population $\mathcal{P}$ consisting of $N < \infty$ population units labelled $u = 1, 2, 3, \ldots, 100$. And suppose that we have a variate's value y_u for each unit u.

Suppose further that we have values of a grouping variate g_u on each unit u which takes the value `"A"` or `"B"` depending on whether u is in subset A or in subset B of $\mathcal{P}$. Here $A \cup B = \mathcal{P}$ and $A \cap B = \emptyset$.

To demonstrate results numerically, define an example population and its variate values for $N = 1000$ as follows

```
set.seed(314159)
N <- 1000
y <- rchisq(N, df = 5)
g <- sample(c("A", "B"), size = N, replace = TRUE)
pop <- 1:N
A <- g == "A"
B <- g == "B"
N_A <- sum(A)
N_B <- sum(B)
data <- data.frame(u = pop, y = y, g = g)
```

Interest lies in the following population attributes and the relations between them:

- the population average (or mean) of the ys:

$$\mu_y = \overline{y}_{\mathcal{P}} = \frac{1}{N} \sum_{u \in \mathcal{P}} y_u$$

- the population variance of the ys:

$$\sigma_y^2 = \frac{1}{N} \sum_{u \in \mathcal{P}} \left(y_u - \mu_y\right)^2 = \frac{1}{N} \sum_{u \in \mathcal{P}} \left(y_u - \overline{y}_{\mathcal{P}}\right)^2$$

  and the population standard deviation σ_y
- the corresponding values for each of sub-populations A and B. That is
  - the sub-population average (or mean) of the ys for $u \in A$:

$$\mu_A = \overline{y}_A = \frac{1}{N_A} \sum_{u \in A} y_u$$

  - the sub-population population variance of the ys for $u \in A$:

$$\sigma_A^2 = \frac{1}{N_A} \sum_{u \in \mathcal{P}} \left(y_u - \overline{y}_A\right)^2$$

    and its sub-population standard deviation σ_A

– the sub-population average (or mean) of the ys for $u \in B$:

$$\mu_B = \overline{y}_B = \frac{1}{N_B} \sum_{u \in B} y_u$$

– the sub-population population variance of the ys for $u \in B$:

$$\sigma_B^2 = \frac{1}{N_B} \sum_{u \in \mathcal{P}} \left(y_u - \overline{y}_B\right)^2$$

and its sub-population standard deviation σ_B

a. *(2 marks)* Write down how μ_y can be determined mathematically from μ_A and μ_B. In R demonstrate this holds for the population values given in `data` above.

Show your code.

b. *(12 marks)* Show mathematically how σ_y^2 can be calculated from σ_A^2 and σ_B^2, the difference in the group averages $(\overline{y}_A - \overline{y}_B)$, and the known group sizes N_A and N_B.

Demonstrate numerically that the derived formula holds by applying it to the population values given in `data` above. Show your code.

c. **Simple random sampling (without replacement).** Suppose we have a sample $\mathcal{S}$ of n units chosen with equal probability and without replacement from the finite population $\mathcal{P}$. Equivalently, there are $\binom{N}{n}$ possible samples of different units u and we choose any one of these samples with probability

$$Pr(\mathcal{S}) = \frac{1}{\binom{N}{n}}.$$

Consider the indicator function

$$I_{\mathcal{S}}(u) = \begin{cases} 1 & \text{if} \quad u \in \mathcal{S} \\ 0 & \text{if} \quad u \notin \mathcal{S} \end{cases}$$

and define $Z_u = I_{\mathcal{S}}(u)$ to be the random variable indicating whether u will be in the randomly chosen sample. Then we can define

$$\pi_u = E(Z_u) = Pr(u \in \mathcal{S})$$

to be the *inclusion probability* of the population unit u, that is the probability that the random sample $\mathcal{S}$ includes the population unit u.

Since $\mathcal{S}$ is randomly selected from the possible samples of $\mathcal{P}$, we can equivalently write

$$\begin{aligned}
\pi_u &= Pr(\mathcal{S} \ni u) \qquad \text{(the probability } \mathcal{S} \text{ contains } u) \\
&= \textstyle\sum_{\mathcal{S} \ni u} Pr(\mathcal{S}) \quad \text{(sum the probability } \mathcal{S} \text{ is selected over all possible samples that contain } u) \\
&= \textstyle\sum_{\mathcal{S} \ni u} \frac{1}{\binom{N}{n}} \\
&= \textstyle\binom{N}{n}^{-1} \sum_{\mathcal{S} \ni u} 1 \\
&= \textstyle\binom{N}{n}^{-1} \binom{N-1}{n-1} \quad \text{(only have to count the number of remaining } n-1 \text{ units in } \mathcal{S}) \\
&= \frac{n}{N}
\end{aligned}$$

The same reasoning can be used to find the *joint inclusion probability* π_{uv} of two different population units u and v. For simple random sampling (without replacement) this is

$$\begin{aligned}\pi_{uv} &= E(Z_u Z_v) \quad (\text{for } u \neq v) \\ &= \tfrac{n(n-1)}{N(N-1)}.\end{aligned}$$

Now, consider the common **sample estimators** of the population average and population variance, namely

$$\tilde{\mu}_y = \frac{1}{n}\sum_{u\in\mathcal{S}} y_u$$

and

$$\tilde{\sigma}^2_{n-1} = \frac{1}{n-1}\sum_{u\in\mathcal{S}}(y_u - \hat{\mu}_y)^2$$

i. *(4 marks)* Prove that $\tilde{\mu}_y$ is **unbiased** for μ_y.

ii. *(10 marks)* Prove that

$$E(\tilde{\sigma}^2_{n-1}) = \frac{1}{N-1}\sum_{u\in\mathcal{P}}(y_u - \hat{\mu}_y)^2$$

and hence that $\tilde{\sigma}^2_{n-1}$ is **biased** for the finite population variance σ^2_y.

iii. *(2 marks)* Show how $\tilde{\sigma}^2_{n-1}$ can be corrected to become **unbiased** for the finite population variance σ^2_y. What happens to this correction as $N \to \infty$?